D-SPIN

D-SPIN Report 2.1: Formation of Centres

June 2009

D-SPIN, BMBF-FKZ: 01UG0801B

Deliverable: R2.1: Formation of Centres

Responsible: Peter Wittenburg

Contents

CLARIN is a research infrastructure project that was selected as one of the ESFRI roadmap projects to be funded by the EC. Its German chapter is called D-SPIN and is being funded by BMBF, the federal states and some research organizations. In this document we will mostly use the term D-SPIN, in many contexts, however, the term CLARIN is meant as well. When we deal with networking experts for example about building an authentication and authorization infrastructure we are dealing with DFN (German national research network) and with eduGain/TERENA (European organizations) at the same time at the two levels. Of course to a large extent D-SPIN needs to adhere to CLARIN agreements, however, D-SPIN partners are actively contributing to these agreements.

1. Introduction

Currently, the Language Resource and Technology (LRT) domain can widely be characterized as an unorganized one in which small accidental networks and temporary collaborations of researchers exist due to projects or individual work. This has led to an utter fragmentation of resources and technology components. The basic goal of D-SPIN is to build a technical infrastructure that helps to overcome the huge integration and interoperability problems, so that it becomes much simpler for humanities and social science researchers to use and combine language resources and technology components.

Centres as Pillars of the Infrastructure

The basis of such a research infrastructure is recognized centres that can offer stable, highly available and persistent services of various types[1]. This is necessary to help to overcome the integration and interoperability problems which exist, and which will continue to arise in a research environment that is characterized by a high innovation rate. The goal is for researchers to be able to concentrate on their research activities, since they are the key for all creative and innovative work. By creating a network of service centres, they will be released from all not primarily scientific tasks. This deal will only work when the individual researchers are not loaded with additional bureaucratic and administrative tasks when accessing services and when the service centres are committed to offer user-friendly services. The integrated domain of service centres to be established offers new possibilities such as working on even larger virtual collections with components gathered and combined from various centres. The new landscape that D-SPIN is aiming at is indicated schematically in figure 1.

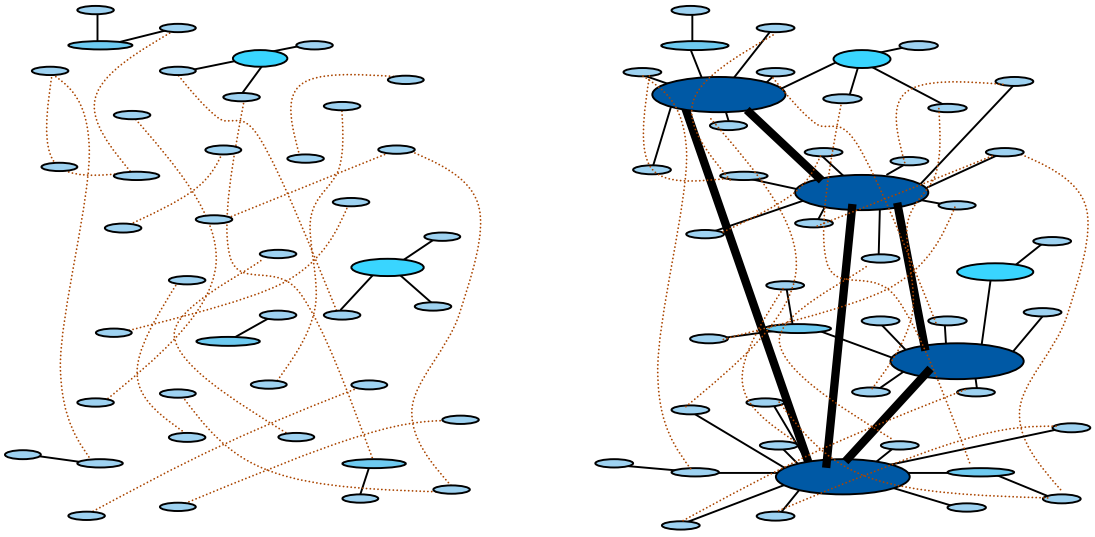


Figure 1 indicates the type of infrastructure change D-SPIN is aiming at. From a scenario characterized mainly by accidental and temporary interactions we want to come to a scenario where dedicated service centres of new type interact in a stable way and give persistent and easy-to-use services to the community. Researchers must be able to rely on the services offered.

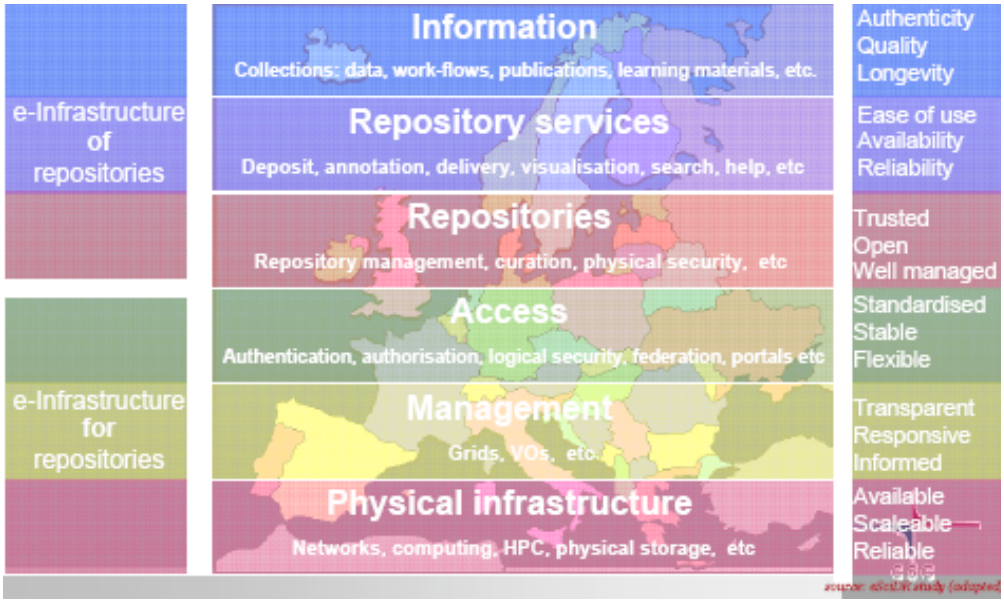
Such a landscape built around service centres is the precondition in D-SPIN's strategy to overcome the fragmentation of resources and technology in the LRT domain. The solutions found in the different countries will vary dependent on the local situations. This includes the question of how many centres there will be in each country. This question cannot be answered; it could even happen that some countries will share centres to reduce costs. In some countries the decision might be to include computer centres or large digital libraries as service hosts, in other countries new types of centres may emerge. In Germany we have a federative governmental structure with a number of strong centres with different foci already now in the various federal

[1] This view is supported by a recent conference of the Alliance for Permanent Access held in Budapest (http://www.alliancepermanentaccess.eu/index.php?id=3) and a recent workshop of e-IRG held in Paris (http://www.e-irg.eu/index.php?option=com_content&task=view&id=106&Itemid=8). Both put the aspect of data management and repositories in the focus of concern and looked at it from different perspectives.

states. The share of tasks will depend on the distribution of skills. In general we can differentiate between a layered system of expertise starting with bit-stream preservation and running powerful servers on the one extreme and giving service that include linguistic knowledge on the other end. D-SPIN will not specify rules how services are implemented, however, D-SPIN will need commitments with respect to the quality and the longevity of a service.

Embedding of Repositories

There is an increasing awareness about the necessity of research infrastructures to take care of the long-term preservation aspects of research data resources. This was indicated at two workshops recently hold[2] and by the document about repositories worked out by ESFRI. Such repositories need to be part of the CLARIN and D-SPIN infrastructure, i.e. there need to be centres that are devoted to the tasks of data curation and preservation.



This figure indicates the scientific data infrastructure according to Kimmo Koski. It shows how repositories and their services are embedded in other layers.

Kimmo Koski [KOSKI] showed the embedding of such repositories within the vertical infrastructure landscape. He is differentiating two layers: the repositories that take care of maintaining the holding and repository services that offer various types of services. Whenever D-SPIN is speaking about data centres we address primarily these two layers although some work needs to be carried out at other layers as well.

There is still an ongoing debate in how far generic research infrastructure services can be separated from those that are domain specific. At the recent e-IRG meeting in Prague this was subject of discussion again, however no clear borderline can be drawn yet. In general it is agreed that we need both the bottom-up approaches driven by the communities and the top-down approaches mostly driven by IT experts who are continuously looking for abstractions.

Benefits for the Researchers

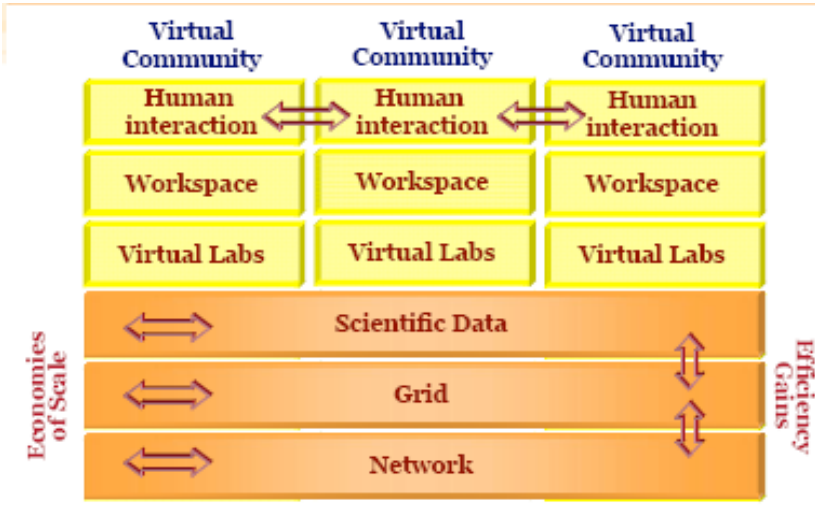
It is known that infrastructures will only be accepted when the participating centres offer real benefits to the researchers. This can be established by a collaborative interaction between researchers and service centres. It will involve:

- researchers hand over (synchronized) copies of their resources and technology to service centres
- in return they will get seamless access without bureaucratic obstacles to a larger variety of LRT via stable and highly available services
- they need to be allowed to integrate resources stored at centres in their teaching work

[2] The APA 2008 Conference, Budapest, http://www.alliancepermanentaccess.eu/index.php?id=3 and e-IRG Workshop, Paris

- they will not be confronted with all sorts of license agreement details which currently hamper innovative work
- they can rely on a proper handling of rights and access restrictions where this is required by legal or ethical considerations, and where they want their own intellectual property rights to be safeguarded
- they can focus on their research work and save much time by easy cross-walking capabilities.

But such centres will also help to overcome a number of other aspects that will only be indicated briefly by one example. The danger of losing our cultural heritage in the digital area is well known and of course language material forms an essential part of our cultural memory, be it as describing cultural events or be it as cultural object itself. According to a study by D. Schüller in the realm of a UNESCO project [Schüller 2004] it turned out that about 80% of our recordings of cultures and languages created in the ethnological/ethno-linguistic domain are highly endangered due to chemical deterioration of the storage media that carry the information. These recordings are stored with individuals in inadequate circumstances. Only when we convince our researchers to hand over such data to centres who know how to digitize recordings and how to migrate to new storage technology at regular intervals we will be able to preserve our material. For much data that is currently stored on notebooks and PCs of the individual researchers the same holds. The material is highly endangered and referring only to research publications without the possibility to go back to the original material would not be satisfactory, since they would only offer a very restricted, interpretive view on data fragments.



This figure taken from Kimmo Koski indicates the way data can be organized on top of basic infrastructure layers and offers a view on how virtual researcher communities can organize themselves, form virtual collections, share workspaces and even allow crosswalks again them. This figure best indicates the evolving benefits for researchers.

At the usage side the diagram of Kimmo Koski [KOSKI] indicates the possible benefits of properly designed infrastructures where we establish a common data layer based on typical grid and federation services and where we allow users to create virtual communities that are building virtual collections and share workspaces.

A similar diagram could be made showing that web services could be combined to new workflows. For operational components, however, the problem of a persistence of service is even more problematic. Researchers need to be sure that a certain tool they want to use will be supported for a while, since in general it costs them a lot of time to acquire the usage skills. Normally, individuals or even research departments cannot give guarantees about persistence, which is leading to frustration and inefficiencies.

Guidelines for Centres

These examples may be sufficient to make the case for carefully selected centres that will be the pillars of the D-SPIN research infrastructure. Some important guidelines need to be followed in selecting, developing and maintaining centres.

1. It is in general better to locate services under the direct control of the persons or groups that created them, since they have deeper knowledge about the content and are more committed. Running a service centre is not merely an administrative task. Long-term persistence of services, and constant innovation and responsiveness to user needs, depend on a deep-seated commitment to the research goals of an academic community. For this reason, data centres located in research centres with an interest in Humanities and Social Science research may be a more appropriate location for LRT type

of services than a large library, for example. But this may differ between country cultures and may change over time.
2. Centres need to provide a variety of services, i.e. they cannot focus just on a particular one, otherwise the work of integration of services across multiple centres may become problematic, and the administrative overhead of running a highly available, up-to-date and sustainable centre is multiplied too many times.
3. On the other hand, for service centres to give realistic guarantees about persistence of services, and to support and maintain them effectively, will mean that they will need to limit themselves to provide a restricted number of services, depending on the resources available to support them.
4. There is also the danger that an important centre has the tendency to establish additional bureaucratic, financial and license hurdles that can hamper research enormously. Keeping the centres in touch with and responsive to the day-to-day concerns of the active researcher is one way to safeguard against the establishment of such additional hurdles. This needs to be separated from the licensing hurdles imposed by the depositors of material for example.

An e-infrastructure can be characterized by three essential ingredients: information and communication technology, information science, and a community. Since research infrastructures are created to serve the researchers' communities and since we should be oriented by user-centred designs, D-SPIN will only be accepted, when we react with a high degree of sensitivity to these concerns and when we are able to mediate between the user interests as resource provider and consumer.

Distributed Infrastructure

The ESFRI process has shown that research infrastructures in the various disciplines are very different. In the Humanities in general, and in the area of LRT (language resources and technology) in particular, what is needed is a distributed infrastructure, not a single location where researchers share a single, very expensive, centralized facility, as may be the case in astronomy, or particle physics, for example. In the LRT domain one can say that researchers in distributed locations will share a distributed virtual facility that is built up on a number of D-SPIN centres.

Yet, we cannot make statements about the number of centres that we will need in a distributed D-SPIN infrastructure, since a number of criteria need to be considered. We can distinguish technological and political/organizational reasons:

- What kinds of services are provided and what kinds of expertise and facilities do they require?
- Which level of redundancy in service provision is required to achieve high availability?
- What are the wishes of the different countries and regions to take over responsibility?
- What can the community do to influence the legal and ethical playground of sharing resources and services?

In particular for economical reasons, it makes sense to follow a strategy of centralization of, for example, server farms and data storage services, as is shown by the experience of some big companies. From the JISC Research Data Digital Preservation Costs Study [Beagrie 2008] it can clearly be seen that there is an economy of scale factor, which states that it is much cheaper for a service centre to offer a larger number of services. D-SPIN, like any other research infrastructure, therefore needs to setup cost-efficient services, but cost efficiency cannot be the only criterion. As in the case of national or regional archives and libraries other important arguments count, such as political and legal issues or the availability of appropriate deep knowledge. Some services can sensibly be centralized - only a few servers are necessary to be setup in Europe to give a registration and resolution service for persistent identifiers (PIDs) for example. A single PID service can easily handle a large number of requests as is shown by [DOI]. It is the need for redundancy and security that will motivate us in such cases to set up a few servers with mirroring functionality. In the case of generic technical services like identifiers, authentication and terminology for example, these can even be shared by other disciplines, if necessary.

Despite technical and economical reasons for centralization in some cases D-SPIN definitively feels the need for distributed services. One of the main reasons certainly is the responsibility for a certain language or dialect spoken in a specific region where we can assume that only regional experts will have the required energy and knowledge to maintain the quality of services at a high level. There is also the notion of a national/regional responsibility for the languages spoken in the country/region and beyond that even a responsibility, partly for historical reasons, for the language material that may come from different regions in the world, but stored in the country.

During the preparatory phase the project D-SPIN needs to come to a good first estimate for the number of centres that need to be set up given the various requirements. Currently about 25 institutions in Europe have identified themselves as being candidates for achieving the centres status. As has been started by the Alliance for Permanent Access [APA] we will also need to describe the costs in detail.

Security Aspects

In many ways distributed services are more difficult to organise than central ones. At the technical level, we need to make sure that there is a continuous communication to synchronize metadata and data, and to ensure conformance across the network to changing procedures and guidelines. For a given PID, for example, all servers need to provide exactly the same resource or resource instances. Therefore it is crucial for a distributed system to rely on secure mechanisms between all participating servers and services to prevent attacks on the consistency of the information for example. D-SPIN needs to take care that appropriate mechanisms are built in from the beginning.

IPR issues will play an important role and need to be dealt with from the beginning. Although D-SPIN wants to achieve a simplification with respect to licenses and, in general, supports the principles of Open Access, we will not be able to define just one model. Open Access needs to be restricted due to personal rights when operating with video sequences for example. The implementation of a "Fair Use" principle with less restrictive rules for researchers would be already very helpful. Therefore, centres need to be prepared to deal with a more complex IPR situation.

Registries

Distributed systems are much more scalable than centralized systems[3], since they exhibit much parallelism and don't require perfect orchestration. However, due to this feature of distributed systems we need mechanisms that can communicate and inform each other about the state of activities. In general, using registration mechanisms for all sorts of activities can solve this problem. The following figure therefore specifies one of the fundamental architecture characteristics of parallel systems where independent service providers offer services to an unknown user community, or, in a web services scenario, to unknown web applications and services. To enable such a scenario to work effectively open registries need to be used that are machine-readable and provide interfaces for human consumption.

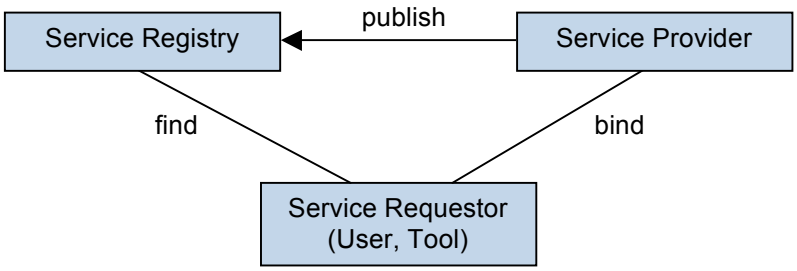


Figure 2 indicates the basic organization principle for a scalable and distributed research infrastructure. Service providers act independently from each other. This requires that all services need to be registered in machine and human readable registries so that users and web agents can find appropriate solutions.

Protocols and Standards

A distributed scenario with a high level of interaction will only operate smoothly when all participating organizations and centres adhere to trust agreements, when they adhere to relevant agreed standards, and exchange information according to agreed protocols. D-SPIN centres therefore need to adopt common standards and best practices where they can be agreed. In those areas where they are missing, or where they are not adequate, we need to push forward new standards. D-SPIN needs to make explicit statements about which standards and best practices will be adopted, and in the construction phase only those centres fully conforming to these can be integrated. In the preparatory phase we cannot impose too strict rules, since most potential centres are far away from adhering to these best practice guidelines.

[3] Also Cloud computing is based on a farm of coupled servers hidden after smart dispatcher software solutions that can easily be extended.

In this document we will not make a detailed list of standards and best practices CLARIN/D-SPIN will adopt. The surveying, analysis and discussion of this area will comprise an important part of the D-SPIN preparatory phase project, and the outcomes cannot be pre-judged. Here, it is sufficient to offer a list of organizations and initiatives that are relevant for our field. In particular we can refer to the following organizations and initiatives:

- Internet and Web: IETF, W3C
- federation and grid: TERENA, GGF, EGEE, DEISA, GEANT
- repository and archiving: ISO (OAIS), OAI
- language resource management: ISO TC37, TEI, IMDI, OLAC

It is the task of other documents to make additional and more detailed specifications.

The Wider Environment

D-SPIN is part of a larger trend to build infrastructures that will pave the way to an e-Science and e-Research scenario, a scenario that has the potential to address integration and interoperability problems with global and sustainable services, allowing the individual researcher to easily access and combine all resources and tools in a simple manner. Despite the high ambitions of the e-Science agenda, it is clear that the realization of these goals is a long way ahead in the future, but D-SPIN is devoted to making significant steps in this direction. This has to be done in constant interaction with other organizations and initiatives, in particular in those areas immediately relevant for the humanities.

Traditionally, libraries and archives are big and important players in providing information to the researchers, as well as to many other constituencies. They have started with large digitization campaigns and are involved in national integration and semantic web projects such as [NESTOR], [CATCH], [TEXTGRID] and the Common Information Environment [CIE]. In addition they are pushing a number of European projects that are of great relevance for D-SPIN. In particular we should mention The European Library [TEL], [DRIVER], [CASPAR] and the Alliance for Permanent Access. DRIVER is establishing an infrastructure where metadata and texts are being harvested in a multilingual context to allow metadata and content searches across repository boundaries. CASPAR is working on data preservation aspects focusing in particular on the "logical level preservation", i.e. all aspects that have to do with data interpretation at levels beyond the bit-stream presentation. It is referring to the results of the EDI infrastructure project [EDI] that focused on appropriate community-based repository structures. APA is looking to increase the awareness about the need to reserve funds and define mechanisms for data curation and preservation to prevent a "dark digital age". Also [e-IRG], the infrastructure reflection group is working on relevant aspects such as proper data management and repository infrastructures. D-SPIN is aware of these initiatives and has already established official interactions as part of its CLARIN activities.

[PLANETS] is in the same family as CASPAR and also offers potential in digital preservation. It might be very fruitful to challenge CASPAR and PLANETS to embed their project results in the context of the ESFRI initiated infrastructure projects, such as D-SPIN and also [DARIAH] and [CESSDA]. These infrastructures will provide a permanent testbed for preservation efforts, and will take digital preservation out of the current subculture approach. With respect to DRIVER, D-SPIN already stated that it will make its metadata registry contents available for OAI-PMH [OAI] based harvesting. In this context also [METS] and MPEG21 DID [DID] are mentioned as container exchange formats for metadata and content aggregations.

Libraries and archives provide a huge amount of material that is of great interest for researchers. Since D-SPIN is focusing on language resources and technology we need to clarify their role and the type of integration we intend. It is the task of an initiative such as DARIAH that wants to organize the humanities as a whole to link up with the libraries and archives. D-SPIN is looking at the libraries and archives from the perspective that they offer a large amount of language resources (e.g. digitized books) that should become available for language oriented operations. Contacts with some libraries at the national level have already been established in Germany such as the SUB Göttingen. A close interaction between DARIAH, the libraries and archives and D-SPIN will ensure that the necessary amount of integration and interoperability will be achieved. In Germany it is intended to join the DARIAH and CLARIN domains in the construction phase.

DARIAH aims to build a network of centres that will include national arts and humanities data centres alongside academic repositories and digital libraries. The scope, encompassing all electronic data types across the arts and humanities, is much broader than that of D-SPIN. One outcome could be that D-SPIN is established as a discipline hub in the network that is being established by DARIAH. D-SPIN centres should become visible within such a network of humanities centres. It is intended to also link up with social sciences

initiatives such as CESSDA, but workflows and terminologies are so different that we can only recommend collaborating at an incidental level. Firm contacts have been established with CESSDA experts and some interaction at working group level has already taken place. These early contacts will be important for defining the interactions between the infrastructures. Of course, there are many more initiatives in the area of the humanities and social sciences and D-SPIN is committed to establish links where possible and in particular to understand how LRT services need to be offered to facilitate the scholars' work.

Research identity federations (IDFs) are emerging in many European countries [NIDF], and also beyond Europe, all aiming to establish domains of trust and to allow their users to act with single identities guaranteed by their home institution. In Germany it is the DFN (German Research Network) that is creating the national identity federation. One of the intentions of IDFs is to make contracts (establishing further trust relations) with service providers such as the big publishers so that researchers can obtain seamless access to resources via their single locally administered identity; many of the addressed research organizations in Germany are already participating in the DFN AAI. [TERENA] and [EduGAIN] are the initiatives at European level that are focusing on a harmonization between the different national IDFs. CLARIN/D-SPIN has established official contacts with TERENA and EduGain at the European level and DFN at the national level.

Yet the idea of having service provider federations is comparatively new. D-SPIN wants to establish such an academic federation providing resources and services in the realm of CLARIN, the basis of which is a limited set of license models, code of conduct rules and limited accounting services. Researchers are both providers and consumers of research publications and datasets, and it is important to separate these roles. At national level one signature will be needed to allow every researcher covered by the national IDF accessing the resources and tools provided by the D-SPIN service provider federation. Therefore, every D-SPIN centre needs to agree on the terms worked out by CLARIN to participate.

Distributed computing at large scale is not a direct priority of D-SPIN. Nevertheless, it is known that for example training large stochastic models that carry out some recognition task on texts (finite state parsers), audio (speech recognition) and video signals (sign language recognition) based on distributed collections high performance computing based on shared file systems will be necessary. Therefore, initiatives such as [DEISA] and [DGI] will be of relevance for D-SPIN at a later moment as well.

With respect to the intentions of D-SPIN we can certainly learn a great deal from cutting edge national initiatives such as those currently pursued by the [JISC] in the UK and D-GRID in Germany. Exploratory studies on various aspects of e-infrastructure, such as architectures for distributed systems, repository setups, registry setups etc. have been carried out, and the experiences will be analysed and used by D-SPIN.

2. LRT Services

The major challenge for D-SPIN is to help the individual researcher who does not have access to a group of experts to build a virtual collection of resources and carry out a large variety of auxiliary operations on the collection. The priority for D-SPIN is to overcome the accessibility, integration and interoperability problems that are currently preventing the researcher from easily or effectively carrying out this work.

D-SPIN will need a variety of centres that can offer stable, reliable and persistent services to make the above scenario a reality. While D-SPIN is focusing on identifying the set of centres in Germany, CLARIN is focusing its activities primarily on building a European network of such service centres, but it will also actively include non-European partners that have already expressed their interest. It is the task of the D-SPIN experts to closely synchronize with the CLARIN activities at the European level. In the following we will describe the types of services that need to be offered in a language resource and technology infrastructure.

2.1. Language Data Resource Services

Language resource services offer online access to electronic language resources of different types such as textual data, speech data, multimodal data or even time series data such as from motion trackers. Each LRT centre in D-SPIN needs to specify which services it will offer under which terms. We can differentiate between a few types of basic services that may coincide in one physical centre:

Advisory services (see also below)

- Language resource archives have found that it is vital to be involved with the creators of resources in the planning stages and during the process of building the resources. This is the most efficient and

cost-effective way to ensure that language resources are created and documented according to adequate standards. This may involve offering an advice service, and ideally also a relationship with funding agencies to ensure that the expertise in language resource archives is used in the planning of resource creation projects, and is available during the resource building process. Language resource centres will also be available to help funders to assess the technical aspects of funding proposals.

- Language resource centres can also offer advice on documentation, conversion and enhancement of existing or legacy resources. In particular, advice about curation principles will be necessary to improve the chances for long-term interpretability.
- Advice on resource discovery is also important, since centres in a federation of language resource archives will have vital knowledge and expertise in where to find relevant resources.

Ingest services

- centres need to allow **uploading of new resources**, which needs to include format checking, format conversion, metadata creation, registration, metadata and content index updating, generating unique and persistent identifiers (PIDs), taking care of versioning, setting the access rights, etc
- uploading revised versions of resources
- optionally also offering the possibility for further user enhancements and additions to resources, allowing repositories to operate as ***Live Archives***[4] in a Web 2.0 scenario

Preservation services

- they need to take care about **long-term data preservation** (archiving and curation). This will involve a strategy for continuous technology migration, and may involve copying the resources and exchanging them with other centres to have multiple copies at several locations. It will include updating the metadata and PID tables; also the associated access services must be maintained

Resource Discovery services

- offering resource discovery metadata in a variety of relevant protocols and standards (e.g. Dublin Core, OLAC, IMDI, TEI, etc.) to enable potential users of the resources to find resources according to a variety of criteria, both by searching or browsing in the centre's catalogue, and also by searching in portals where metadata from the various centres in the federation are aggregated. Resource discovery metadata needs to be available for both humans (end users) and machines (aggregators and language processing components).

Access services

- they need to allow users to **access data** (both metadata and resources) via standardized APIs and suitable web interfaces to support access by both machines and humans; to ensure this a centre needs to be a full member of the LRT federation (see note 2); both metadata and content access methods need to be included: fast searching via indexes and browsing/inspection; a wide variety of access methods is expected that allows to access collections, individual resources as well as fragments of resources even if this is as fine-grained as the value of one attribute in a lexical entry.

A number of requirements that such centres need to fulfil, such as persistent access to uniquely identified objects, follows immediately from the mentioned types of basic services. Other requirements are for example: (1) They need to offer a large amount of storage capacity. (2) Its services should not be limited to national contributions, but need to be offered at European level and for all disciplines that want to deposit language resources. (3) They need to have a suitable and reliable repository system and eventually an archiving strategy.

Centres, of course, can delegate services and still be part of a trusted domain as long as the delegation is formal enough and therefore can support the commitments made.

2.2. Language Technology Services

Other types of centres will offer tools that can be executed on resources and resource types. Again the type of tools will include a broad spectrum such as tools working on texts, on speech signals, on video signals to extract multimodal behaviour and many more. Each LRT centre in D-SPIN needs to specify which services it will offer under which terms. Also here we can distinguish a few types of services: (1) Creating services; (2)

[4] http://www.mpi.nl/dam-lr/lra-flyer/

finding services; (3) computing services and (4) advice about using the complex software. The latter service is summarized below under infrastructure services, here only the special requirements for the first three services are sketched.

(1) There are many tools around and many of them need to be made real web services so that people and computer programmes can make use of them. This implies that the tool code needs to be encapsulated to become a D-SPIN compliant service and that the tool needs to be registered in a searchable and browsable registry. The registry will include pointers to the interface description. This procedure of making tools available to the D-SPIN user community is not that simple, since encapsulating existing software requires mostly a decent level of expertise. Often small research groups will not be able to carry out the encapsulation themselves; therefore D-SPIN service centres specialized on this task need to help these research groups

(2) Similar to resources it must be possible to search and browse for useful services in the registry. There need to be processing options for humans and machines such as for example an automatic profile matching. The registry entry will contain a metadata description, a manual about its usage, a formal specification about the input/output formats, some information about run-time aspects, an API specification (REST[5], WSDL) and PIDs pointing to a specific service in a unique way.

(3) Running computationally intensive jobs on large datasets entails logistical challenges, i.e. the data has to be brought to the computers that execute the job or in some cases it may even be much simpler to transport and install the code to carry out a job. The HTTP protocol used in normal web services interaction is comparatively slow and not suitable for high performance computing. In the DEISA project tests were made with optimized distributed file systems to solve the problem of providing data from remote locations fast enough to keep parallel computers busy. In CLARIN it is intended to test such a work scenario between a few selected institutes, but this work is not in the primary focus. Nevertheless, it could be possible that certain centres offer specific services such as conversion or annotation services. The general idea is that a remote application uses an API (REST or WSDL) offered by a service provider, submits datasets via HTTP and will receive a data set back via HTTP. Such services require adequate descriptive metadata, registration with an appropriate discovery mechanism for such tools, and a persistent, available interface as indicated above.

An application suggestion service needs to be delivered by the infrastructure providers. When a user has selected a resource of a certain type he should get suggestions for suitable programs to carry out a specific task. The INTERA project proposed the Language Repository Exchange Protocol [LREP] to handle such requests. Using LREP a request with a resource profile was sent from the resource provider to the software registry and a list with a set of possible tools was returned. This is an important service, but the early LREP solution needs to be adapted to the actual requirements.

Summarizing, we can state that there will be centres offering access to computational tasks. These centres will convert applications developed by research groups to executable web services so that data provided by the user (or calling service) can be processed, albeit via the existing, relatively slow channels. In most cases resource centres will themselves offer a variety of access and analysis tools. All these technology components require descriptive metadata and registration in the D-SPIN tools registry as well.

2.3. Infrastructure Services

Beyond these two fundamental service types about resources and technology components we need services that guarantee the smooth functioning of the infrastructure as a whole. Often these services are core services and their high availability is of crucial importance for the whole infrastructure. The necessary robustness and high availability is likely to require that they be served by a number of centres to create redundancy. Yet it is not obvious in how far these services can be used by other disciplines as well or whether D-SPIN can make use of services offered by others.

For example, four such services are described below.

Metadata Services

[5] For REST-full services there is no widely used agreements about the interface specification. WSDL and WADL have been suggested.

It is obvious that one of the core tasks within D-SPIN will be to setup an agreed infrastructure for a registry mechanism for all resources, tools and perhaps even advice. A human and machine readable registry infrastructure is essential for creating a domain where all language resource and technology components are registered, described with metadata and where pointers direct to the services. The LRT community has a long history of metadata standards and their application. IMDI was defined in 2000 by a European group of linguists, OLAC was an extension and refinement of the Dublin Core set, and [TEI] provides a descriptive framework for some main data types, and has been used by a number of projects. In particular, [IMDI] and [OLAC] developed an infrastructure that allowed to search across repositories. IMDI already introduced the notion of a linked infrastructure of metadata descriptions which comes close to what is required for the registry infrastructure and it allowed users to create temporary virtual collections. Also the interoperability between both domains, IMDI and OLAC, was solved by introducing gateways.

D-SPIN needs to be built on a comprehensive and flexible registry setup fulfilling several needs that became apparent during the last 8 years. Such a service needs to be based on a redundant set of service centres that offer all functionality. German service centres need to take care that all relevant national services are registered and that the German contributions can be harvested to establish centres with European (and even beyond) scope. The details of the metadata infrastructure for D-SPIN have been worked out in the CLARIN Component Metadata Infrastructure [CMDI] document.

PID Services

It is widely accepted that URLs are not adequate to point to stable and persistent services. Therefore, various suggestions were made to overcome the unstable situation. Recently, a proposal was made to ISO to introduce unique and persistent identifiers (PIDs) for language resources and services. The proposal was widely accepted at the recent ISO TC37 meeting in Provo [TC37]. The Max Planck Society followed these suggestions as a whole and in the mean time a consortium is being formed covering large computer centres that will offer a robust and highly performant PID registration and resolution service based on the Handle System. In Germany the GWDG in Göttingen will offer the service to researchers in particular from D-SPIN but probably also to a wider research community. D-SPIN should strongly support and push forward the ISO standardization and should advocate this service, making clear however that centres can use any other robust PID resolution system that is highly available.

Again, we can state that a D-SPIN infrastructure will necessarily need to support PIDs. The details of the PID service aspect have been dealt with in the PID document [PID].

Schema and Concept/Terminology Services

It has become widely understood that the main pillar of interoperability is the creation of a domain of shared or interrelated concepts. Nevertheless, it is good for many reasons such as harmonization and standard compliance to offer generic models for certain resource types as well. This approach has been chosen for the Lexical Markup Framework [LMF] as being specified by ISO TC37/SC4, for example. Using shared concepts registered in a data category registry will be the way to set up a new more flexible metadata infrastructure in CLARIN and the component paradigm has been adopted by TEI as well. This implies that users can build their own schemas and integrate them to new ones as long as they make use of registered and predefined categories. To prevent a proliferation of schemas in such a liberal scenario it must not only be possible to create such a schema, but to register them, to allow others to re-use them etc. Therefore, we will need well-known schema registry services in CLARIN. Since such a service will be contacted whenever parsing of a schema is required the services need to have a high availability, thus requiring redundancy.

As indicated ISO TC37 recently made the step to tackle the semantic interoperability problem. It defined the ISO Data Category Registry standard (ISO 12620) [DCR] and already now we can see that the ISO DCR is being filled with many important concepts used in linguistics so that they will become available in the near future for re-usage and reference purposes. However, it was well understood that this DCR will just be one pillar in a more complex world that will emerge existing of a number of centres that will offer the central ISOcat[6] concept service to achieve a high availability, probably a few other centres with alternative terminologies and a number of centres offering the possibility to register relations between registered concepts and between such concepts and tags used for example in schemas. It is obvious that D-SPIN will need to take up these issues, discuss them in detail, suggest further solutions to ISO where necessary, add

[6] ISOcat is the concrete service that implements the ISO DCR standard and offers appropriate services to the users. It will be maintained under the CLARIN initiative.

categories used by German linguistic community such as the Stuttgart-Tübingen Tag Set for example and take care of appropriate services.

The D-SPIN preparatory phase project will need to evaluate the potential requirements for these services, and make recommendations for appropriate resource allocations in the construction phase. While schema, concept and relation services are of undoubted value in computational linguistics and computer science, it is not yet clear to what extent this concept registry infrastructure can be used by users in the Humanities to create and maintain practical ontologies to be used for semantic weaving for example. Yet no widely agreed and applicable methods have been implemented.

Advisory Services
Centres giving advice will form important pillars in the D-SPIN network. We have to foresee a wide number of different types of advisory services and yet it is too early to create a comprehensive list and to define strategies of how to tackle all aspects. Here we can only indicate the large variation of needed advice:

- advice on sustainability and the potential for the reuse of data
- resource formats and linguistic encoding wrt. all linguistic data types
- application of standards for texts and media
- usage of tools and availability of methods
- repository and archiving advice
- interoperability methods and conversion possibilities
- multilingualism experts
- usage of registry and registration services
- availability of experts about all aspects
- tool installation
- usage of APIs
- training people to maintain the infrastructure
- advisory services to funders (advising on and assessing technical aspects of research grant applications)

It needs to be sorted out when all these aspects will be taken up, which strategy will be chosen to accomplish this and who will take responsibility. Some of the activities should be taken up rather soon; others need more time to plan. Nevertheless, it is obvious that we need a registry for advisory services and that people and institutions listed need to take over responsibility on a European level. It is the task of another work package to set up an index for these advisory services.

3. LRT centres

3.1 Centre Types

Based on a deep analysis of the kind of services D-SPIN will offer and the goals of D-SPIN to establish a stable and persistent research infrastructure where researchers can rely on the services we have defined 5 types of centres and describe them in appendix A in more detail. In this context we only want to briefly indicate the various types:

- Recognized centres (Type R) offer language resources or tools, but their offering is not integrated into the D-SPIN infrastructure except by referring to a web site and a contact person.
- Metadata providing centres (Type C) offer machine readable metadata that can be harvested via the accepted protocols so that users can browse and search in the D-SPIN registries to find these resources and services.
- Service providing centres (Type B) offer services that include the access to resources and tools via specified interfaces that are D-SPIN compliant.
- Infrastructure centres (Type A) offer services that are relevant for the functioning of the infrastructure as a whole and that need to be given with a high level of commitment.
- External centres (Type E) that offer D-SPIN relevant services but which are not from the list of D-SPIN members.

The status of each centre can change over time and it is obvious that in particular in the preparatory phase some of the D-SPIN member institutions will not have the funds to participate at a more advanced technological level. D-SPIN does not want to exclude any group that has excellent resources, but we also

want to stress that it is the purpose of D-SPIN to convince researchers and projects to hand over their resources to centres of their choice to make them available for users of the infrastructure,

In the preparatory phase we will ask all members what kind of services they will be able to offer at a certain time. At the web site there will be a description of each centre that will be used as the core for building up a machine-readable registry of centres.

3.2 Business Models for Centres

The business model has to support the typical work pattern of modern research not only in the LRT area, but also in particular in the humanities and beyond:

- access patterns to resources are not predictable, they are dependent on the actual insights, questions which are changing dynamically and on sub-discipline specific views on the data;
- often quick inspections on large virtual collections of resources originating from different resources are of highest relevance to see whether certain resources may contain answers to the questions in mind;
- in the same way a toolkit, comprising a set of different tools, is required to allow users to carry out a variety of operations on the selected resources in a simple way; users want to play and test, since often they are not sure whether a certain sequence of operations will lead to the answer they are looking for;
- simple-to-establish workflows have to help researchers to overcome the interoperability problems, i.e., if a useful tool cannot be carried out immediately on a resource help should be given to transform the resource;
- unlike data in some scientific domains, data in the humanities and social sciences remains of interest to historical studies after it ceases to be representative of contemporary states of the language, and any synchronic dataset can potentially be of use for historical and diachronic studies – data does not go out of date. Therefore long-term preservation and continuous access are important factors to consider.

These requirements mean that it is necessary for centres to offer standardised, interoperable, persistent and sustainable services to users.

In the LRT domain some centres have already been established at the European and the German level, such as ELDA and IDS. The ELDA "business model", however, is mainly suitable for supporting industry that want to get full control of a certain data collection and therefore are willing to spend money for such a complete resource collection. Some specific research projects may require this type of access as well.

Other types of centres, such as the Oxford Text Archive, offer a wide variety of free resources for download and use, and offer a range of ingest services and options, but do not offer deep access to the content of resources, or possibilities for cross-searching and combining resources, and do not currently offer web services. These features have been implemented by digital repositories/archives such as the one at the DOBES/MPI. Centres as IDS are reorganizing their services to even adapt more closely to the research process described above.

Some existing centres do offer options for sophisticated queries and processing, but only to a limited set of resources and with limited views. D-SPIN needs to overcome many of these restrictions. To allow all researchers to participate and to foster highly innovative research a new type of service centre and new business models need to be implemented by D-SPIN, facilitating the provision of resources and tools as interoperable and persistent services.

At a recent workshop organized by the Alliance of Permanent Access statements were made about possible business models in the general sense for preserving research data and it was basically concluded that we can speak about an imperfect market since the value of research data cannot be estimated, i.e. only public funding will work. The same holds for research infrastructures in which repositories will be embedded. Some cost estimates were made for repositories indicating that there are a number of unknown factors. Making cost estimates for infrastructures is even more problematic at this moment.

4. General Architecture

In this document we do not want to present an abstract overall architectural diagram, but discuss architectural requirements that have implications for D-SPIN centres along a few dimensions and derive a number of basic principles that will guide the D-SPIN technical infrastructure work. A general architecture for centres that participate in the UK Infrastructure project was given by A. Powel (see figure 3). It describes some basic functional blocks a repository system should include along the three dimensions content, middleware and information services. For D-SPIN we can refer to this functional schema, but need to go a step further.

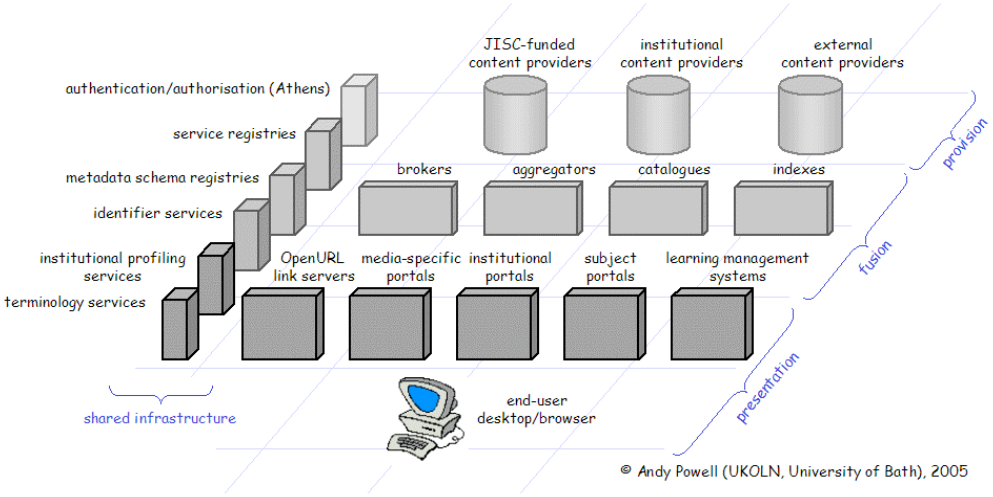


Figure 3 shows the JISC Information Architecture as worked out by Andy Powell.

4.1 Principles for D-SPIN's Distributed System

CLARIN has already now 32 participating countries and we can assume that each country will have at least one dedicated centre to participate in the network. In several cases such as in Germany we will have several centres with specializations. This means that CLARIN and D-SPIN will establish truly distributed systems and we can also assume that configuration despite all declarations of funding support will vary over time. In figure 2 we already gave a schematic impression about the type of loose interactions we intend to build. Here we first want to describe a few guiding principles.

Principle of Independence: Every participating centre is independent in its choices of internal organization and setup as long as it adheres to the agreements that are defined for a smooth interaction within the network.
It is obvious that most of the potential centres have a history where local infrastructures emerged over time to offer some services. No one can and will expect that these local infrastructures will completely be replaced if they can be adapted to match with the requirements.

Principle of Service: Every participating centre needs to make an explicit statement about the services it wants to give and about the quality characteristics of these services.
The distributed infrastructure needs to be able to rely on certain offers, since otherwise no promises can be made with respect to the users. This does not include per se statements about the quality of the content.

Principle of Consistency: Every participating centre needs to guarantee that the content it provides when a unique and persistent identifier is used to refer to the content will not change over time.
Users need to rely on the fact that references are stable and lead to the content they have selected.

Principle of Interoperation: Every participating centre needs to adhere to the set of interaction protocols and agreements defined within D-SPIN/CLARIN.
Protocol and agreement conformity is required to guarantee a smooth interaction of locally diverging architectures. Protocols include specifications about the formats of information exchange.

Principle of Responsibility: Every participating centre takes over a responsibility for the coverage of the services it offers.

In a large distributed infrastructure hierarchies of responsibility need to be established to cater for manageability. In such a layered system each service layer has to rely on the proper operation of the other layers.

4.2 Entities in Distributed Scenarios

In highly distributed scenarios as envisaged by D-SPIN we should describe the entities that are participating. That perspective will help to formulate additional requirements for D-SPIN centres.

Researchers as Providers
Researchers and researcher groups will continue to provide data and algorithmic resources. They have the expectation that they can deposit data resources and ask someone to help with integrating algorithmic resources into the D-SPIN service infrastructure. They need to provide metadata and may want to associate some attributes (rights etc) with the resources and these may change over time. The integration services must be simple since the researcher may not be confronted with unnecessary overhead.

Researchers as Users
Researchers want a flexible view on the resources that are offered within the infrastructure, create their own virtual work environments by selecting and combining resources, carry out enrichments of various sorts such as annotations and relation drawing and execute operations on the selected resources. They expect that services are available that support cross-collection and cross-archival work and that they can act with their academic identity that is granted by an appropriate institution. In the context of D-SPIN we will often use the word "researcher" as user, but in many occasions this will include other employees and members of academic institution such as students.

Researcher Groups
Researchers often act in smaller or larger groups and they expect that an infrastructure will support collaborative work. This can include even so-called virtual organizations.

Services
The basic function of an infrastructure is to offer a wide range of easy to access services (see chapter 2). It is important that they can be found, are offered persistently and can interact with each other. Metadata, protocols and service quality statements take care that this is guaranteed. Services in general have several instances to provide a high availability. It needs to be ensured that these instances generate the same results.

Centres
Centres will offer these services by adhering to the fundamental principles and requirements. Centres will take care of secure interaction channels.

Data Resources
Data resources are the atomic objects that can be addressed in the infrastructure with the help of a PID or URI. It is the responsibility of the providers to define the granularity of resources, to provide metadata and to make suggestions for an embedding in a canonical organization where necessary. In general it is expected that resources be described with a schema that is openly accessible. Data resources are accessible via services that guarantee that the user will always get the same content when specifying a unique ID.

Collections
Collections are bundles of data resources that are stored in repositories and that can be associated with a unique and persistent identifier, own metadata and attributes. Collections can be created crossing archive boundaries. Collections are virtual in so far as they are represented by metadata descriptions that refer to the real resources, i.e. no resources are copied or moved when metadata descriptions are exchanged.

Applications/Services/Tools
Tools are operational resources that are available with atomic or aggregated functionality. In the web they exist as web applications and web services, the latter just having program interfaces to let them be invoked by other applications or services. It is the responsibility of the providers to define their granularity, to provide proper metadata descriptions, a formal description of the interfaces and functionality and to deliver code according to the D-SPIN/CLARIN specifications. Also with respect to applications there must be information about the versions so that it is possible to verify what has been changed.

Workflows
Tools can be combined to create new more complex operations if the import/export specifications match. Workflow frameworks need to ensure that users can combine tools and the needed data resources graphically and that they can save the execution context.

Registries
Registries are the most important data resources to be handled in an infrastructure, since they are the glue that has the potential to bring resources of different types virtually together. Basically they include machine and human readable metadata descriptions and where necessary pointers to interface specifications and other attributes needed in executable settings. Domains where registries will be needed are: data resources, web applications and services, virtual collections, workflow contexts, resource schemas, linguistic concepts, relations between concepts, servers, persistent identifiers, resource responsibility, resource aggregation and distribution, user profiles, etc. Registries must be openly readable and be specified by schemas.

Protocols
Also protocols are essential for a smoothly functioning distributed system, since they specify how independent servers and services can interact with each other. Many protocols are given by different standardisation organizations. D-SPIN/CLARIN will check where specifications are missing and make suggestions. Partly the protocol specifications also specify the exchange formats to be used.

API Specifications
APIs specify a set of methods that a service offers so that it can be invoked by other applications or services. It also specifies the kind of parameters such as the nature of the input and output formats to be provided. For web services specifications have been worked out by W3C and within the Grid community.

User Interfaces
User Interfaces specify the presentation formats for human consumption, i.e. how information is presented to the human user and how the user can interact with an application. Increasingly often these user interfaces are user customizable, so that a view on the information is given that is optimally suited for the user. User wishes need to be registered in user profiles.

4.3 Federation Elements

The term "federation" was introduced into the area of distributed information and communication infrastructures to describe domains of trust and security based on a number of agreements. In DAM-LR and a few other projects the term "federation" was analyzed in more detail. In further working groups these aspects will be worked out for D-SPIN in more detail. Here we just want to introduce the basic terms that are required to identify requirements for centres that want to participate in the D-SPIN language resource providing federation.

Distributed Authentication and Authorization
In Europe and other areas a number of national identity federations have been built and new ones are in the state of establishing themselves. Basically they specify the kind of attributes that are required to identify a user or a class of users, how user management needs to be done so that everyone within and outside of the federation can rely on the correctness of the attributes and how the attributes can be exchanged via secure communication lines. To achieve a harmonization at European level the TERENA and EduGain projects have been started in the realm of the GEANT network project.

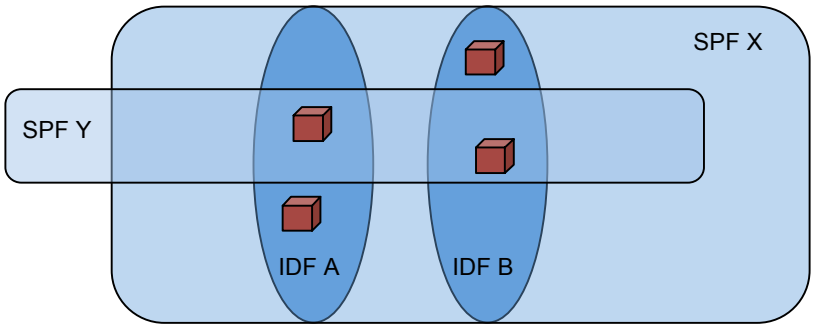


Figure 4 shows a possible scenario where two identity federations established a domain of clear rules about user management. One service provider (federation) obviously includes all members of the two IDFs. Another SP(F) just includes only some of the IDF members, however, they can rely on the same rules.

Identity federations (IDF) are making contracts with service providers. Partly, the academic institutions themselves appear as content providers such as in the case of education material. IDFs also sign contracts with external service providers such as the big publishers of electronic publications. The advantages are that it just takes one signature to allow all researchers who are covered by the IDF to access the corresponding material and that a researcher can access all material based on one identity - the one that is given him by his home institution. Federations don't solve the problem of access permissions where this can only be granted to selected persons included in an identity federation. At the side of the resource provider appropriate records need to be created. But given that the user has access permissions the mechanisms described here are the basis for creating virtual collections crossing repository boundaries.

Also service providers can form federations if they specify the set of attributes they rely on, how they expect attributes being used and what the license conditions are, that may include financial, legal and ethical statements. Since there will be a wide range of different service providers and since IDF members will not make use of all offers member institutions of an IDF will be part of a variety of such federations covering IDFs and SPFs (see figure 4).

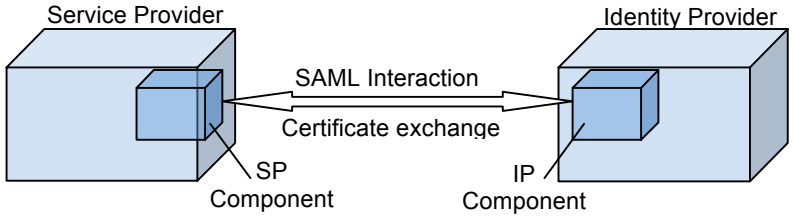


Figure 5 indicates the interaction between service and identity provider and the type of interaction that is required.

Currently mechanisms such as IP-based or proxy-based identification are in place, but it is known that they offer many restrictions and management problems. To overcome these severe disadvantages middleware has been developed and protocols have been specified to take care of a smooth and secure interaction between the participating institutions, the federations are just mediators to establish a ground for mutual trust. SAML 2.0 is now widely accepted as protocol for exchanging user attribute information; so also D-SPIN centres will need to adhere to this specification (see figure 5). Of course the servers that interact in this scenario have to identify themselves in a secure manner, each centre should have their certificates signed by root certificates taken from the TERENA/TACAR list [TACAR].

Currently two middleware components have been tested and installed: (1) The Shibboleth [SHIB] software developed within the Internet2 project can be seen as mature and it provides components for the IP and the SP sides. (2) SimpleSAMLphp [SimpleSAML] is a component for the identity provider side, created by FEIDE [FEIDE] and especially being tested in some Nordic countries.

Part of a large distributed AA infrastructure such as planned in CLARIN is the identification of the centres that are part of it. As is shown in the Chinese university museum project [Tansley 2006] for example it makes sense to maintain a registry of these centres with a number of attributes that can be used by infrastructure services. For most of such registries important for the functioning of the infrastructure we intend to establish domains of responsibility, redundancy and aggregations the principles of which are indicated in figure 6.

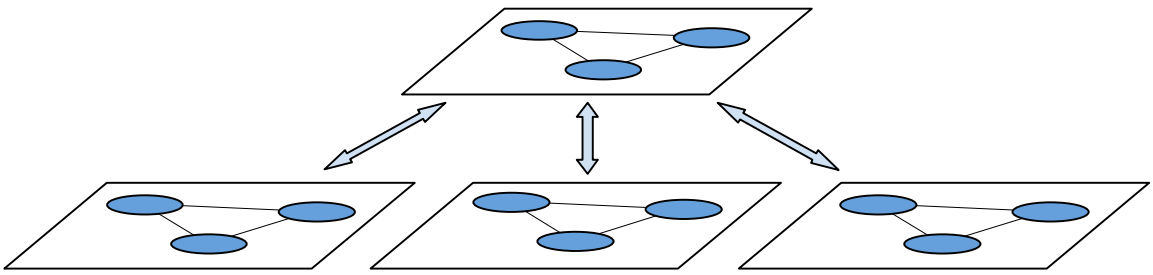


Figure 6 indicates levels of responsibility in the aggregation and management of infrastructure information. Here information about centres participating in the service provider federation is aggregated similar to DNS systems. At European level a few redundant services are collecting the information from the national levels. The national services are responsible for their domain. Not in all cases redundancy at national level will be necessary.

A German coordination centre needs to ensure that the German centres provide valid information and carry out the national aggregation of information. At European level centres acting as supra-national service providers need to do the additional step of aggregation. Each level needs to set up portals that allow humans to select and visualize information. It will depend on the type of service and the national situation how much redundancy will be necessary at the various levels. In general D-SPIN needs to work out requirements to allow a dynamic exchange of information so that at least always one of the services is operational.

Persistent and Unique Identifiers
Another middleware service to be offered in a distributed infrastructure in which resources will be exchanged and enrichment services will be offered is the registration and resolution of persistent and unique identifiers. It is obvious that only persistent and unique identifiers (PIDs) will guarantee a survival of the many references that will be established. In a CLARIN document [PID] a detailed analysis of various suggestions and systems was given. Yet there are different opinions about the question whether well-chosen HTTP URIs have a better chance to survive compared to explicitly assigned numbers that are maintained in explicitly setup distributed databases. In the realm of D-SPIN a PID registration and resolution system based on the Handle System has been set up to allow the introduction of PIDs for infrastructure resources where necessary.

Not all centres participating in D-SPIN will be ready to associate PIDs in the form of Handles with their resources, some may want to use DOIs, others that are closely linked with national libraries may want to rely on URNs and some may rely indeed on well-chosen HTTP URIs. The Handle System based registration and resolution service will be open for all CLARIN members to register their resources if needed. In addition D-SPIN needs to work out a strategy that allows dealing with various types of PIDs, however, it will need to rely on the persistency guarantees specified by the centres. Given certain heterogeneity, D-SPIN will probably need a way to make formal statements about the resolution of PIDs per centre.

4.4 LRT Registries

Since data and tool registries (DTR) with rich metadata descriptions will be dealt with in particular by special working groups we will restrict this discussion to a few basic aspects. DTR is a key concept in distributed infrastructures, since they are the domain to register all known resources and services so that users and machines can find them. For human users it is like a catalogue of a big data warehouse where products from a large variety of producers will be made available for re-usage. For machines it is the source of information to automatically find appropriate services to solve a problem.

Metadata descriptions can be seen as the incarnation of the object they describe, they include relevant information identifying an object in form of keyword-value pairs. For a number of operations they can be used instead of the resources themselves such as for searching on metadata categories or for building virtual collections. Metadata can be used for a wide range of purposes such as management of large collections, discovery, selection, citation, virtual collection building and direct research purposes. Since they act as easy to use fingerprints of the resources themselves, repositories and archives must ensure that the link between metadata and resources is carefully maintained, i.e. in general one will assume that metadata is delivered with a PID as reference to the resource. Thus as the most general case we can design an architecture as denoted in figure 7.

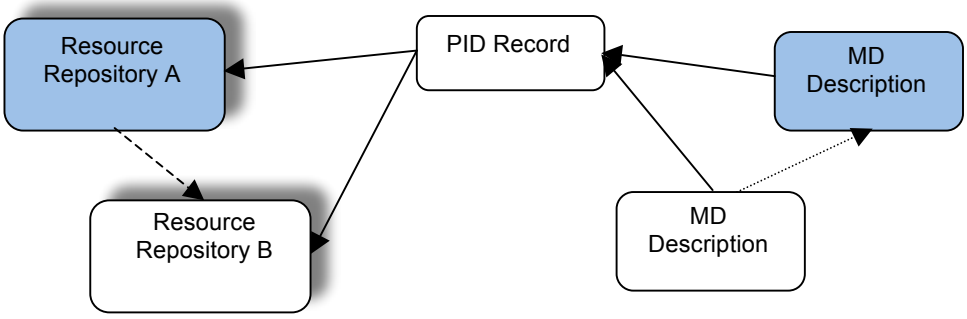


Figure 7 indicates the central role of persistent identifiers in accessing resources. Resources will be copied for several reasons and the PID record needs to point to all instances of a resource. Also metadata descriptions that refer to the resource objects can be copied in many different virtual collection specifications. Therefore it makes sense that the metadata descriptions point to the PID records.

There can be several instances of a resource for various reasons such as redundancy to take care of data preservation and there can be various copies of the metadata description, since it is open and everyone can harvest it and use it for some purpose. The unifying resource is the PID record that is unique, i.e. all MD descriptions will point to the PID record and the PID record will contain pointers to the different resource instances. It is the task of the resource providers to take care that changes to the original resource are propagated to the various locations where instances are stored, i.e. basically a push concept needs to be implemented. It is the task of the metadata harvesting sites to take care that they have up to date metadata descriptions, i.e. basically a pull concept is being implemented.

As was already indicated the PID service is crucial, i.e. redundant services being synchronized dynamically need to take care that a high availability is achieved. With respect to metadata registries a layered network of responsibilities is of great importance in CLARIN/D-SPIN. There need to be dedicated portals that aggregate and organize metadata at the German and at the European level and thus can react quickly if a centre shows problems with respect to the commitments made. This canonical and well-maintained infrastructure is important, since it operates in a domain where everyone can create arbitrary linked structures of metadata descriptions, i.e. there has to be one that will be always available as reference structure.

Metadata is open and can therefore be harvested by any service provider. Currently, two harvesting methods are well known: XML resource and OAI-PMH [PMH] based. Any accessible XML file can be harvested and be interpreted if the underlying schema is well described and if the vocabulary used adheres to agreements. Another method is to use the widely known OAI-PMH protocol to request and exchange metadata. In addition to the underlying schema a Dublin Core [DC] based record needs to be delivered, i.e. a gateway performing some form of schema and vocabulary translation is required.

Therefore, it is required in D-SPIN for all centres offering resources in making their schemas explicit, refer to the vocabulary specified in accepted data category registries and offer appropriate metadata descriptions. Some will support the OAI PMH protocol, but this is not yet required. However, for the service providers at the German and European level we require that they provide a DC mapping and an OAI PMH port. When harvesting, the canonical structure where provided and harvesting attributes should be stored to allow offering a browsable domain according to the original categorizations of the creators. In particular for metadata it holds that the national levels are responsible for maintaining a proper registry of all contributions from the various centres.

It is left to further documents to make more detailed statements.

4.5 Formats

Many different types of data resources such as (structured) texts, annotations, lexicons, images, video signals, audio signals, time series data (eye tracking etc), tables, protocol messages etc will be stored and exchanged in D-SPIN. These are contained in various formats according to encoding, structuring and packaging principles. In this chapter we want to briefly clarify some vocabulary and again define requirements for D-SPIN centres.

Each data resource has user or machine relevant data, which we may call its content. The content is encoded according to documented principles such as a well-structured text containing sequences of UNICODE characters encoded in UTF-8 that are structured according to an XML schema for example. For storage and other operating system purposes this storage is packaged in some way that is not relevant to the user. We call this the representation format. It is the task of the repository system to generate from this internal representation format the same content independent of time and circumstances. Archives are responsible to preserve this content for long periods, which is called the principle of bit stream preservation.

Content can be stored in various containers such as relational database systems. Mostly these systems apply some form of data encapsulation to achieve more optimal representations and access conditions. Additional software is required to generate the content information again that adheres to the encoding specifications. Many repositories are using such containers. D-SPIN will not make any specifications, but will require that the repositories are capable to generate the content on request.

The user may want to visualize content and can use various options such as printers and displays. It is the task of application software to present the data in some form. Users will also make selections and apply

filtering to restrict the amount of content to be presented. These methods are often described with the term "view". Here many preferences and user customization options play a role so that D-SPIN will not make any statements except that the user needs to be ensured that the basic information generated is the same.

The cyber-infrastructure opportunities extend to various forms of enrichments. The extension of existing collections of resources by new ones is uncritical since new objects will be added and integrated into the repository. More often researchers want to add commentaries, draw relations between fragments and add new layers of annotations or lexical attributes for example. These are operations that directly refer to existing content and closely related to the structure of the data. D-SPIN repositories need to make sure that such extensions do not change the content of the existing objects. Either new object versions need to be created or the extensions need to be stored in other resources together with the references.

Versioning is an important issue to be dealt with by repositories. It must be possible for the user to retrieve the same content again. In general this can be ensured by identifying each version with a unique identifier. However, there are repositories that are subject of many and frequent changes such as for example a database that stores commentaries to a great number of different resources. The repository needs to make explicit statements about its version policy and the way it wants to ensure that users can rely to get the content they expect.

Increasingly often exchange protocols that come with format specifications are required. The purpose is to specify the packaging of content in a standardized way to make its interpretation independent of any special packaging within the repository. Even container formats are increasingly often defined to allow the exchange of complex information including various resources, their metadata and even their relations. We can refer here to standards such as SOAP, METS, MPEG21 DID, OAI PMH and SRW/SRU [SRW, SRU]. D-SPIN will make strong statements about the type of protocols and information packaging that it will use and D-SPIN repositories are expected to adhere to these standards.

4.6 Services

Complementary to data resources D-SPIN will offer a large number of services. Also here it is left to future documents to go into more detail. This document has the purpose to clarify some aspects that are relevant for D-SPIN centres.

First, we would like to separate between local applications running on individual PCs (applications), applications that can be executed over the web (web applications), services that can be invoked by other services or web applications in the web (web services) and web sites which just present some information for human consumption.

D-SPIN will mainly focus on web application (WA) and web services (WS) that are offered to run in the infrastructure so that users (and machines) can execute them, i.e. tools that only function locally may be registered and described by metadata to make them visible. D-SPIN will need to specify the way WA and WS will be registered, described by metadata and invoked. It will rely on existing standards where possible.

It is the task of centres to take care that WA and WS are offered in a persistent way and in such a way that the user knows what he can expect. Since software is often subject of changes we cannot expect that older versions will be maintained, but for each such service an explicit statement has to be maintained about the changes. This area is still subject of research in computer science so that further statements will be left to later documents.

5. Repository System

It has been widely accepted that proper repository or archiving software systems are the core facilities for centres that want to offer data resources in a stable and standardized manner and that want to participate in networks or federations of interacting centres. As already indicated each centre in the D-SPIN language resource and technology federation should be free to choose their own system. But discussing a few of such systems may help to define criteria, to adapt the architecture and perhaps to come to new decisions.

We need to make a difference between serving data resources or processing resources (execution of some application or service software), since for running web applications or services often special run-time environments are required asking for specifically setup computers etc. In this chapter we only discuss typical software packages that can help to manage large amount of data resources, to add new resources in a

structured way, to maintain metadata about the resources, to distribute data for various purposes and to integrate the resources into a federation.

5.1 Topics

To start we first would like to discuss a few basic topics that emerged from various discussions.

Repository/Archive/Digital Library
The term "repository" describes an entity that stores, manages, offers and distributes data of all sorts in a serial, atomic and neutral way, guaranteeing that users (or machines) will get the content they are asking for. Actually they "de-reference" PIDs in the form of URIs and Handles. The term "digital archive" describes a repository that has a clearly specified long-term preservation strategy. The term Digital Library emerged from the domain of libraries and their extension towards "digital libraries". DLs have a bias towards the typical publication type of resources, but DLs want to address a broader scope and its description includes a variety of services that are offered to access and combine the resources. In this document we have a preference to use the terms repository/archive, since in the LRT domain often special services such as search engines and presentation tools are offered together with the resources that emerged from our discipline.

Live Archives
The term "Live Archives" has been defined within the [DAM-LR] project to describe the fact that digital research repositories/archives are living entities in so far as researchers want to enrich and modify the existing collections and to combine them in various ways to new virtual collections. Since continuous, but unpredictable access patterns are essential any form of restriction would hamper research. Of course all changes and enrichments may not affect the existing resources, but must be included in a standoff manner.

OAIS Reference Model
The OAIS model (see figure 8) is widely accepted as a sort of reference model to check the appropriateness of a given repository architecture. It is based on packages of various sorts. Central is an "Archival Information Package" that contains the content to be stored and the associated Preservation Description Information. AIPs need to be administered carefully where descriptive information (metadata) plays an important role. The resource producers provide "Submission Information Packages" that should include all relevant creation information. During ingesting the data the archive will add own information to create the descriptive information and to create the complete AIP. At the consumer side a number of typical services are offered to find and access resources. Finally the data is delivered as "Distribution Information Package".

Several projects used the OAIS model to check proper archiving, but it is widely known that it is underspecified for this purpose. It can also not cope with the ideas of Live Archiving in all respects. Nevertheless it is a very useful reference model.

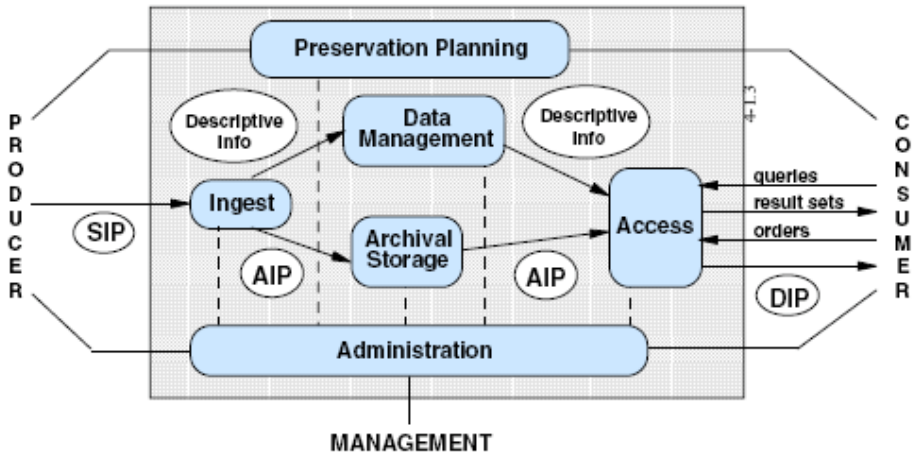


Figure 8 shows the well-known OAIS model which is widely being accepted as a reference model for building digital archives. In particular it gives a good view about the need to distinguish data management which is based on descriptive metadata and archival storage

Data Seal
Due to the deficits of the OAIS model a number of initiatives started with working out "checklists" that allow archive managers to assess the appropriateness of their repository/archive architecture and rules of operation. The rational behind all these efforts is to take care that the data we store can be found, accessed and interpreted after a period of time. We can refer here to activities from the Digital Curation Centre [DCC],

[NESTOR], [DRAMBORA], [TRAC] and [DANS] which all focus on the issue that a set of criteria should be fulfilled to accept a centre as being a proper repository/archive. D-SPIN should adopt these methods and ask the participating data resource centres to do a self-assessment at regular intervals, since finally our users who want to store their data should be able to rely on the promises.

Container Formats
Container formats have been specified to bundle content, their metadata descriptions and their relations amongst each other. The idea of container formats originates from the OAIS model, since it is seen as utterly useful to store all information that belongs together in one context. The interpretation is, however, very heterogeneous. Some see container formats as means to exchange data in proper described schemas to make the interpretation independent from repository internal formats. Some see containers in accordance with OAIS as the basic units of storing resources with long-term preservation intentions. Some just bundle metadata descriptions since it does not make sense to include lengthy multimedia resources in such a container. The inclusion of relations and references is not at all clear since most of them point to resources outside of the resource bundle. A number of initiatives made suggestions such as METS, MPEG21 DID and now more recently [ORE]. Repository systems such as [FEDORA] are built explicitly on an object model that is similar to the ideas of container formats. D-SPIN will not make any specifications about the internal storage format; however, it will need to come to a specification about the interchange format. Here one of the widely used formats will be supported and D-SPIN centres need to provide this format.

Encapsulation
Archives often require storing data in a way they are deposited, i.e. if an XML file has been uploaded, the file should exist as XML file and should be identifiable as such at the level of the operating system or file system. This is frequently seen as important to guarantee interpretation of the file without additional software. Several software systems such as relational databases or XML databases are encapsulating the resources, i.e. when uploading they turn the content into an optimal internal representation. Only additional software thus will allow the user to find, access and interpret the data. Mostly the packages guarantee to deliver at export the same data as it was deposited, but this claim needs to be checked. D-SPIN will not make any prescriptions in this respect, i.e. as long as the repository can guarantee that all promised resources are available persistently and unchanged D-SPIN will be satisfied.

5.2 Solutions

A number of solutions is being discussed and has been developed with different foci. It needs to be mentioned that repository systems in general do not include application-oriented solutions. The logic to present a lexicon at the user interface and to allow all operations lexicographers would like to have etc is in general not part of the software packages offered and needs to be developed on top of the existing repository code. It cannot always be seen whether the logic pre-fabricated into a given system will hamper certain applications or not. It is known for example that accessing FEDORA objects or XML files is not very fast, i.e. for certain operations different intermediate formats need to be created.

File Systems
Some institutes just use existing file systems to store resources and use a web server such as Apache to give access to them. In general, there is a number of missing functionalities of such systems that needs to be provided by separate software. For example no metadata support is given, since the directory structure only gives the path of nodes associated with short names. In general a file system alone will not be sufficient, although there are currently some projects with the intention to associate more information with nodes.

Relational Database Systems
Many institutions are using rDBMS from commercial vendors (ORACLE, DB2) or from the open source market (mySQL, Postgres) as a container to include all kinds of linguistic information. Due to the relational approach there is no problem to establish a table structure that contains for example resources (even video or structured texts), annotations on those resources, metadata descriptions, relations between all resources, administrative information about access rights etc. All resources are encapsulated, i.e. internally they are stored in some binary format and one has a dependence of the rDBMS software.

Some apply a mixed strategy and use an rDBMS for all layered information and include references for media files for example, which are not stored in the database. These packages all come along with an SQL based access and management language and allow developing any kind of application logic on top of the system. In the D-SPIN case certain applications need to be provided to ensure that for example metadata harvesting can

be done and that data authenticity is granted. For long-term data preservation these systems cannot be recommended, since the interpretation is dependent on additional software.

Some of the professional systems such as ORACLE come with full-blown solutions to build distributed databases, but the solutions are all included in the proprietary code. D-SPIN would not make itself dependent on a company in this respect.

XML Database Systems
Similar statements can be made for good XML Database Systems such as eXist for example. Such databases are specialized to host XML-structured data, but also apply some kind of internal encapsulation. The approach can be compared with rDBMS although professional rDBMS still have a development advantage in some respects such as performance for high data volumes.

CMS
Increasingly often content management systems are applied to manage and access web-based content. Here also various proprietary and open source solutions (PLONE/ZOPE [PLONE], [DRUPAL] etc.) are on the market. They started as systems to maintain a large number of web pages, but emerged to full-blown content systems that also support programming languages such as Python to develop and include application logic. The more general of them support a wide range of standard document types such as XML, PDF and Word Doc and they do not apply per se apply encapsulation.

FEDORA
Fedora provides sustainable technologies to create, manage, publish, share and preserve digital content. It's a system that is based on an object oriented container model that allows combining content, metadata, relations and methods operating on the content. It provides a set of APIs to manage, access and manipulate the FEDORA objects. However, the set of available functions if rather limited compared to SRB for example. Therefore, much functionality needs to be added to create a full-blown repository system. FEDORA expects that a broad user community will create additional functionality that can be re-used by others, dependent on the task serious development work is required.

eSciDoc
[eScidoc] is new development of the Max Planck Digital Library based on the FEDORA object model to make it useful for storing, managing and accessing all kinds of data. Currently, an alpha version of a publication manager has been developed and a few application packages show that eScidoc can deal with photos and other objects as well. The focus of the work was on managing publications coming from all Max Planck Institutes. The amount of developed code is much greater than the code provided by FEDORA, indicating the effort that is needed to make FEDORA usable. eScidoc is certainly a very interesting development, since it includes relevant features such as metadata, persistent identifiers and AAI components.

SRB/iRODS
Storage Resource Broker [SRB] is a data grid middleware software system produced by the San Diego Supercomputer centre (SDSC) that combines repository and data exchange/distribution functionality. Depending on the "flavour" of the configuration, use patterns, and policies, the SRB creates what is called a data grid, a digital library, persistent archive, and/or distributed file system. SRB provides a uniform interface to heterogeneous data storage resources over a network. As part of this, it implements a logical namespace and maintains metadata on data-objects, users, groups, resources, collections, and other items in an SRB Metadata Catalog stored in a relational database management system. System and user-defined metadata can be queried to locate files based on attributes as well as by name.

The SDSC SRB system is middleware in the sense that it is built on top of other major software packages such as database software and it has a callable library of functions that can be utilized by higher-level software. Thus, SRB can work with various types of sources, does not include an encapsulation and is very flexible. In comparison to other packages it comes along with features for distributed data management and allows users for example creating virtual data collections.

Recently, the same group announced iRODS, which is based on SRB, but in addition offers an engine that interprets rules to decide how the system is to respond to various requests and conditions.

SRB presupposes the existence of SRB clients on all participating instances that cannot be assumed in D-SPIN. SRB certainly is one of the most powerful solutions for distributed data storage and management. All application logic needs to be created by the community of course.

LAMUS

[LAMUS] is a comprehensive repository system developed by the MPI for Psycholinguistics with a clear focus on language resources and a number of typical applications for manipulating lexica and multimedia annotations. It allows configuring which data types and formats are accepted and one can associate parsers that check format correctness, in particular XML is supported. All data including metadata is stored in open file formats, i.e. no encapsulation is done and all relevant content can easily be extracted or copied to another repository that is independent of LAMUS. As basis for its organization and management the IMDI metadata infrastructure is used. It provides functionality to efficiently deal with large amounts of resources, to efficiently set access rights, to associate resources with persistent identifiers, to integrate it into AAI based federations. Other tools can be added easily on top of the existing ones.

DSpace

[DSPACE] is software that provides a way to manage research materials and publications in a professionally maintained repository to give them greater visibility and accessibility over time. It emerged from needs from libraries and institutional repositories to manage in particular publications. DSpace captures text, video, audio, and data, offers distribution and indexes features allowing users to search and retrieve items. DSpace comes with a number of pre-fabricated applications and has a more restricted data model compared to FEDORA. It is widely used in Digital Libraries all over the world.

ePrints

[ePrints] is a software system developed at the University of Southampton in the UK for the ingest, storage, archiving, description, discovery and distribution of scholarly electronic outputs. While it was originally conceived as a repository system for publications, it now claims to function for repositories of research literature, scientific data, student theses, project reports, multimedia artefacts, teaching materials, scholarly collections, digitised records, exhibitions and performances. ePrints is currently in version 3, and claims to be "the world's most popular repository software"[7] It is oriented towards the needs of UK higher education institutions, many of which are currently engaged in setting up institutional repositories.

Google/Flickr/etc.

Increasingly often general users make use of "free" web offers and store their data with companies like Google, Flickr, Yahoo, etc. This seems to be a very attractive option, since full indexing is offered allowing full-text search on the texts stored. One can assume also that at least the big companies will offer robust and fast access to the resources via their standard interfaces. Due to various reasons we cannot recommend to create a dependency on a company's business policy for scientific data. It is completely open in how far D-SPIN would get access to the stored data to allow combining it with others.

D-SPIN Policy

As already indicated D-SPIN will not make any statements with respect to choosing one of these systems as long as they allow machine access to the resources or their metadata descriptions via clearly defined interfaces and as long as centres can guarantee that a PID will be resolved into exactly the same resource or service independent of time.

6. D-SPIN Requirements

In this chapter we will briefly summarize the requirements that finally should be fulfilled by institutes that want to participate as centres in the emerging CLARIN/D-SPIN infrastructure.

- Centres need to offer useful services to the D-SPIN community, which can be a mixture of metadata descriptions, data resources, processing components and/or infrastructure services.
- Centres need to agree with the basic principles:
    - Independence, i.e. free choice of internal organization of service
    - Service, i.e. explicit statements about services, the duration and the quality
    - Consistency, i.e. guarantee to deliver the same content for the same identifier
    - Interoperation, i.e. adhere to agreed protocols and agreements

[7] http://www.eprints.org/software/

    - Responsibility; i.e. commitment to take over responsibility as part of D-SPIN
- In general D-SPIN will expect that if a service once offered to D-SPIN will be stopped the service will be transferred to another centre to guarantee continuity for the user community.
- Centres need to adhere to all security guidelines, join a national IDF where available and be ready to join the D-SPIN SPF.
- Repositories in D-SPIN should have a proper and clearly specified repository system and participate in a quality self-assessment procedure as proposed by DANS or others.
- Each centre needs to make clear statements about their business model and their treatment of IPR issues. In D-SPIN OPEN ACCESS and fair use principles need to be supported were possible.

7. D-SPIN Centres Network

7.1 Procedure

CLARIN has created a template that allowed those institutions that would like to participate as centres within its emerging network to carry out a self-assessment. This self-assessment serves a number of functions:

1. Each institute needs to make explicit statements about its technological and funding support state and its perspectives in these respects.
2. Every self-assessment will be subject of an interaction with the leader of the technical infrastructure work to clarify the statements and to improve and unify them.
3. The TI leader commented the state descriptions to indicate what needs to be done to fully comply with the evolving D-SPIN requirements.
4. An interaction with the institutes resulted in a time planning for each potential centre.
5. A formal call for participation in January 2009 determined which institutes want to play a role as early adopter in the D-SPIN preparatory phase. Also the type of centre was specified.
6. All information was bundled, commented and prepared so that decisions could be taken which institutes will be integrated in which role and what time at which conditions.
7. These decisions were made public and turned into activity.

This process is not closed, i.e. other institutes can indicate their intentions to become a centre by filling in the appropriate forms and in doing so start the process.

7.2 Candidate Selection

Currently, the following institutes have passed the process successfully and are busy to adapt their strategies:

- Berlin-Brandenburgische Akademie der Wissenschaften, Berlin — B centre
- Institut für Deutsche Sprache, Mannheim — A centre
- Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen — B centre
- Deutsches Forschungs-Zentraum für Künstliche Intelligenz, Saarbrücken — B centre
- Universität Leipzig, Fakultät für Informatik, Leipzig — B centre
- Max Planck Institut, Nijmegen[8] — A centre

These institutes are currently busy in adapting their strategies to fulfil the requirements. The result of the interviews and their major offerings can be seen via the CLARIN web-site:

Interview results: http://www.clarin.eu/wp2-documents/wg-21-documents
Resource & Tool offering: http://www.clarin.eu/inventory

Due to the different states in the reorganization process, four of these centres have been selected to participate in the very first trust federation which will be realized in 2009 including also centres from the Netherlands and Finland: BBAW, IDS, DFKI, MPI. After having successfully set up this test federation, other candidate centres will be added.

In the mean time further institutes have indicated their interest to play a role as centre in future:

- Universität Hamburg, SFB Mehrsprachigkeit, Hamburg
- Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren, Trier
- Bayerisches Sprach Archiv, München

[8] MPI will offer its services to all German linguists such as its archiving service and further push infrastructure services in collaboration with other institutes such as GWDG.

First discussions have taken place in June/July 2009 and we expect that the assessment process will be started and finished in 2009.

Yet for all candidate centres it is obvious that written commitment statements from the responsible funding agencies and an indication of sufficient capacities will be required to participate as a centre in the construction phase.

8. References

Projects and abbreviations

[APA]	Alliance for Permanent Access	http://www.alliancepermanentaccess.eu/
[CASPAR]	Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval	http://www.casparpreserves.eu/
[CATCH]	Continuous Access To Cultural Heritage	http://www.nwo.nl/nwohome.nsf/pages/NWOP_5XSKYG_Eng
[CESSDA]	Council of European Social Science Data Archives	http://www.cessda.org/
[CIE]	common information environment	http://www.common-info.org.uk/
[CMDI]	Component Metadata Infrastructure	http://www.clarin.eu/files/wg2-4-metadata-doc-v5.pdf
[DAM-LR]	Distributed Access Management for Language Resources	http://www.dam-lr.eu/
[DANS]	Data Archiving and Networked Services	http://www.dans.knaw.nl
[DARIAH]	Digital Research Infrastructure for the Arts and Humanities	http://www.dariah.eu/
[DC]	Dublin Core	http://dublincore.org/
[DCR]	Data Category Registry	http://www.isocat.org/
[DEISA]	Distributed European Infrastructure for Supercomputing Applications	http://www.deisa.eu/
[DID]	Digital Item Declaration	http://en.wikipedia.org/wiki/Digital_Item
[DOI]	Digital Object Identifier	http://www.doi.org/
[DRAMBORA]	Digital Repository Audit Method Based on Risk Assessment	http://www.repositoryaudit.eu/
[DRIVER]	Digital Repository Infrastructure Vision for European Research	http://www.driver-repository.eu/
[DRUPAL]		http://drupal.org/
[DSPACE]		http://www.dspace.org/
[EDI]	Electronic Data Interchange	http://en.wikipedia.org/wiki/Electronic_Data_Interchange
[EDUGAIN]	GÉANT Authentication and Authorisation Infrastructure	http://www.edugain.org/
[EGEE]	Enabling Grids for E-sciencE	http://www.eu-egee.org/
[e-IRG]	e-Infrastructure Reflection Group	http://www.e-irg.eu/
[ePrints]		http://en.wikipedia.org/wiki/Eprints
[eSciDoc]		http://www.escidoc-project.de
[FEDORA]	Flexible Extensible Digital Object Repository Architecture	http://www.fedora-commons.org/
[IMDI]	ISLE Meta Data Initiative	http://www.mpi.nl/IMDI/
[ISOTC37]		http://en.wikipedia.org/wiki/ISO/TC37
[JISC]	Joint Information Systems Committee	http://www.jisc.ac.uk/
[LAMUS]	Language Archive Management and Upload System	http://www.lat-mpi.eu/tools/lamus
[LMF]	Lexical Markup Framework	http://en.wikipedia.org/wiki/Lexical_Markup_Framework

[LREP]	Language Repository Exchange Protocol	See [Broeder 2002]
[METS]	Metadata Encoding and Transmission Standard	http://en.wikipedia.org/wiki/METS
[NESTOR]	Network of Expertise in Long-Term Storage of Digital Resources	
[NIDF]		http://wiki.rediris.es/tf-emc2/index.php/Federations
[OAIS]	Open Archival Information System	http://en.wikipedia.org/wiki/OAIS
[OLAC]	Open Language Archives Community	http://www.language-archives.org/
[ORE]	Object Reuse and Exchange	http://www.openarchives.org/ore/
[PID]	Persistent Identifier Requirements	http://www.clarin.eu/files/wg2-2-pid-doc-v4.pdf
[PILIN]	Persistent Identifier Linking Infrastructure	https://www.pilin.net.au/
[PLANETS]	Preservation and Long-term Access through Networked Services	http://www.planets-project.eu/
[PLONE]		http://plone.org/
[PMH]	Protocol for Metadata Harvesting	http://www.openarchives.org/OAI/openarchivesprotocol.html
[SCHU]		
[SHIB]	Shibboleth	http://shibboleth.internet2.edu/
[SimpleSAML]	SimpleSAMLphp	http://rnd.feide.no/simplesamlphp
[SRB]	Storage Resource Broker	http://en.wikipedia.org/wiki/Storage_Resource_Broker
[SRU]	Search/Retrieve via URL	http://www.loc.gov/standards/sru/
[SRW]	Search/Retrieve Web Service	http://en.wikipedia.org/wiki/Search/Retrieve_Web_Service
[TACAR]	TERENA Academic CA Repository	http://www.tacar.org/
[TEI]	Text Encoding Initiative	http://www.tei-c.org
[TEL]	The European Library	http://www.theeuropeanlibrary.org/
[TERENA]	Trans-European Research and Education Networking Association	http://www.terena.org/
[TEXTGRID]		http://www.textgrid.de/
[TRAC]	Trustworthy Repositories Audit & Certification	http://www.crl.edu/PDF/trac.pdf

Literature

[Beagrie 2008]	Beagrie, N., Chruszcz, J., and Lavoie, B. (2008). Keeping research data safe.	http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf
[Broeder 2002]	Broeder, D., Wittenburg, P., Declerck, T., and Romary, L. (2002). LREP: A Language Repository Exchange Protocol. In Proceedings of the LREC 2002 Conference. Las Palmas, May.	http://www.mpi.nl/IMDI/documents/2002%20LREC/LREP%20%20A%20Language%20Repository%20Exchange%20Protocol.pdf
[Schüller 2004]	Schüller, D. (2004). Safeguarding the documentary heritage of cultural and linguistic diversity. Linguistic Archive Newsletter, 1(3):9-10.	http://www.mpi.nl/lan/issues/lan_03.pdf
[Tansley 2006]	Tansley, R. (2006). Building a Distributed, Standards-based Repository Federation. D-Lib Magazine , 12 (7/8), 1082-9873	http://dx.doi.org/10.1045/july2006-tansley

Appendix A: Centre Types

1. Introduction

The D-SPIN infrastructure will not be a monolithic institution offering all of the Language Resources and Technologies (LRTs) for use in the Humanities and Social Sciences. D-SPIN recognizes that the creation, deployment, development and support of LRTs is not restricted to members of the consortium, or even to members of the wider network. The structure of D-SPIN must allow innovation and variety in resource creation and provision, and that users have a choice of services and can select the relevant one for a particular task. The D-SPIN infrastructure therefore needs to have a distributed architecture, allowing many different levels of involvement in infrastructure activities.

This document describes five classes of centres within D-SPIN that are necessary to build and operate the infrastructure. This also takes into account that the transition to a full-fledged D-SPIN LRT Federation will be a gradual, step by step process reaching out to the construction phase, since the funding for centres will come from various sources, and will therefore come at different times.

It is important to note that being a D-SPIN 'centre' in the senses that are defined here is not the only way to be involved in D-SPIN. Many vital roles in the preparatory phase and beyond will be played by centres of expertise in the creation, development and use of LRTs. Furthermore, many vital roles in surveying and research, the development of guidelines and standards, training, dissemination, administration, etc. will continue to be played by key members of the D-SPIN network, who will not necessarily be service centres in the infrastructure.

Therefore, not all D-SPIN member institutions will be D-SPIN centres in the senses defined here. The criteria for classification as a centre at certain levels defined below are more appropriate for centres to be service-oriented institutions such as a language resources repository, a Grid computing centre or other computing service. Other activities are more easily carried out by research institutions or academies. Yet many of the most important centres of expertise in the D-SPIN consortium are university departments, which are more oriented towards fixed-term research projects, and these projects and activities tend to be reliant on particular individuals. Academic departments are by no means excluded from centre status, but they will have to pay special attention to the issues of business models, funding and sustainability. Furthermore, academic departments and other types of institutions will continue to play key roles in projects, provide representatives to boards, working groups and committees, and work in the development of standards and guidelines etc. Being a centre is by no means a precondition for D-SPIN-related funding, or participation in activities, however, from a centre we expect a certain degree of commitment to deliver specified services over a long period.

The preparatory phase in particular requires the input of the individuals and institutions with the highest levels of expertise and understanding of the research practices. Many participants will play a role in developing services that will be deployed in other centres outside of their own institutions.

2 Overview

We distinguish between five classes of centres:

- Infrastructure centres - Type A Centres
- Service centres - Type B Centres
- Metadata Centres - Type C Centres
- Respected Centres - Type R Centres
- External Centres - Type E Centres

In a short table we can summarize the differences except for external centres:

	R	C	B	A
Online services	X	X	X	X (opt)
Services and harvestable metadata accessible via a CLARIN portal		X	X	X (opt)

Fully integrated CLARIN-conformant services			X	X (opt)
Core, essential infrastructure Services with service level definitions				X

The last three levels of service need to be associated with high availability and long-term commitments, which means typically a statement of support until 2020.

During the lifetime of D-SPIN centres can change their type in both directions. When commitments can no longer be given the state will change, however, this needs to be notified to the Executive Board early enough. There should always the option be used that services are transferred to another institution to keep long-term availability in the focus. In case of infrastructure services this would be a must as well as when new commitments can be made. Also the requirements for these centre types will change over time, since new insights in the pillars of an infrastructure will come up. Also in this respect centres may change their type, but formal discussions with the Executive Board are required. For the D-SPIN centres we will introduce a candidate status during the preparatory phase taking care of the fact that it will take a while until the requirements are met.

The categorisation of centres should not be seen purely in evaluative terms, such that A is better than B, which is better than C, etc., and in teleological terms, such that all institutions participating in D-SPIN must strive to move up the ladder towards A status. This is not necessarily desirable because the most efficient and effective way to organise the infrastructure is not likely to be with only type A centres. An ecosystem of centres at the various levels is much more likely to provide the necessary variety and flexibility that users will require, and that will be necessary for continuous evolution and enhancement.

On the other hand, we do need to build up an impressive constellation of services, and for appropriate institutions, there will be rewards in terms of prestige and allocation of resources for those who have the willingness and the ability to build effective centres as backbones of the infrastructure to be built. D-SPIN needs to create an environment where success in building and enhancing centres is rewarded, but where there are also rewards for other types of activity.

The following detailed description of centre types aims to help institutions to define the type of centre that they might wish to become in the D-SPIN infrastructure.

3. Detailed Description

We expect from all centres registered[9] in D-SPIN that they offer at least resources and/or tools or that they control access and grant usage rights to such resources if they are for example deposited at another centre.

3.1 Type R Centres / Respected Centres

For Type R centres it would be sufficient, if these resources and tools will be explained and offered via traditional web sites. No further requirements are defined except the registration of the web-address and a contact person. However, there should be an interest to become a centre actively participating in the D-SPIN LRT Federation.

3.2 Type C Centres / Metadata Providing Centres

In addition these centres offer their services via harvestable metadata and have a portal where the access ways are explained.

- Typically a human- and machine-readable catalogue and a schema describing the structure of the metadata descriptions is provided.
- To support harvesting either the base address for OAI PMH based harvesting or the base address for XML harvesting is provided.
- It must be possible to access the resources from the metadata description and there need to be clear statements how access to the resources can be achieved.
- If a processing component is offered, users should be able to access it via a web site.

[9] At the Web-Site there will be a prominent place which shows the registered centres, the type of services they are offering and the nature of the commitment statement.

3.3 Type B Centres / Service Centres

In addition these centres are full members of the D-SPIN LRT Federation; however, they still act as individual centres not taking over responsibilities for the federation. There needs to be a clear long-term commitment for their services.

For **data resource** centres we can describe the following criteria:

- Their resources will be maintained in a well-structured and documented repository system with a long-term commitment. Versioning will guarantee that references will remain valid.
- The repository system is associated with an accepted PID[10] (persistent and unique ID) service (either in the institute or by making use of registrations at another accepted PID service site). The institute takes care that resolving the PIDs will lead to the correct object.
- The repository system will interact with a Shibboleth resource provider instance to participate in the distributed D-SPIN AAI federation. For its own users it will either be member of a national/organizational identity federation or will setup its own SAML2.0 based identity provider and link it with its local authentication system.
- The centre offers various access paths to the resources amongst which there is a structured access to the complete resource for those who will need this type of unfiltered access.
- The access to the resources of the centre will comply with the license templates outlined in D-SPIN.

For **processing components** offered by centres we can describe the following criteria:

- There needs to be a well-structured and documented architecture in which the services are embedded.
- The processing component needs to be augmented by web services specified by WSDL or REST descriptions indicating the available methods which are accessible via the metadata descriptions for example.
- There needs to be a clear statement about accessibility and service quality.
- The centres need to support a SAML2.0 based resource provider instance to allow users to access the services via their home identity. For its own users it will either be member of a national/organizational identity federation or will setup its own SAML2.0 based identity provider and link it with its local authentication system.
- The access to the services of the centre will comply with the license templates outlined in D-SPIN.

3.4 Type A Centres / Infrastructure Centres

These centres take over responsibility for helping to manage the federation in one way or another. These centres normally extend their services described under Type B Centres; however, these centres could also be computer centres, for example, taking over some service of high relevance within D-SPIN such as running a PID registration service open for all. There is a whole range of possible roles such centres can take over such as:

- running a national metadata registry, harvesting/maintaining national providers, offering a specific portal and providing a OAI PMH gateway for central harvesting
- running a (national) PID registration service
- offering large scale computing facilities to the community to execute compute-intensive tasks
- offering virtual collection services allowing to integrate collections from different sites
- offering workflow service to specify and execute chains of operations from different sites
- offering central services for the AAI federation
- etc

3.5 Type E Centres / External Centres

A further category of centre forming part of the complete infrastructure is proposed, to cover centres offering key services which are necessary to the successful operation of the D-SPIN infrastructure or centres that are from different domains such as a national libraries etc. An example of the first would be a national certification authority. Numerous D-SPIN centres will be reliant on these certification centres for dealing with the bulk of access, authorisation and authentication requests, but the certification authorities themselves will not be D-SPIN centres.

[10] PIDs can be either URIs or Handles of which the persistence is guaranteed.

Looking out from D-SPIN more widely, there will also be related infrastructures and services which form a part of the research infrastructure environment, of which D-SPIN is a part. These are categorized as Type E centres, since they may provide services of which D-SPIN centres and users make use. For example, a sociolinguist may make use of social science numeric datasets accessed via the CESSDA infrastructure, and may wish to combine this data with linguistic datasets acquired via D-SPIN. It is anticipated that D-SPIN will work towards appropriate levels of interoperability with other infrastructures to enable successful resource discovery, access and processing, but CESSDA will remain outside of the D-SPIN infrastructure.

To give another example, a repository in another continent might offer D-SPIN-conformant services, such as resource discovery metadata, or text processing tools, and could thus usefully be categorized as an External Service Provider.

The E Type of centres will give us a possibility to also list them formally in the centres registry.